CoToHiLi at LSCDiscovery: the Role of Linguistic Features in Predicting Semantic Change

Ana Sabina Uban[♠,♡] Alina Maria Cristea[♡] Anca Dinu[♣,♡]
Liviu P. Dinu[♠,♡] Simona Georgescu[♣,♡] Laurențiu Zoicaș [♣,♡]

[♡]Human Languages Technologies Research Center, University of Bucharest
[♠] Faculty of Mathematics and Computer Science, University of Bucharest
[♣]Faculty of Foreign Languages and Literatures, University of Bucharest

auban@fmi.unibuc.ro, alina.cristea@fmi.unibuc.ro, anca.dinu@lls.unibuc.ro
ldinu@fmi.unibuc.ro, simona.georgescu@lls.unibuc.ro, laurentiu.zoicas@lls.unibuc.ro

Abstract

This paper presents the contributions of the CoToHiLi team for the LSCDiscovery shared task on semantic change in the Spanish language. We participated in both tasks (graded discovery and binary change, including sense gain and sense loss) and proposed models based on word embedding distances combined with hand-crafted linguistic features, including polysemy, number of neological synonyms, and relation to cognates in English. We find that using linguistically informed features combined using weights assigned manually by experts leads to promising results.

1 Introduction

In recent years, more and more studies in computational linguistics have focused on the issue of lexical semantic change, tracking the shift in the meaning of words by looking at their usage across time in corpora dating from different time periods (Hamilton et al., 2016; Schlechtweg et al., 2020). Vector spaces and word embeddings have widely been used for tracking semantic shifts of words across different time periods.

Previous studies on the computational analysis of lexical semantic change have found that different word properties such as word frequency and polysemy have a role in influencing the potential semantic shift of the word, proposing statistical laws of semantic change such as the law of innovation and the law of differentiation (Hamilton et al., 2016; Xu and Kemp, 2015; Uban et al., 2021b, 2019). Uban et al. (2021a, 2019) have proposed that semantic change can be studied cross-lingually, by comparing present meanings of cognate words, which by definition share a common etymon from which the current meanings have diverged. The resulting implication is that analyzing cognates of the target word in other languages can also potentially provide clues regarding the word's prior semantic change. We provide more details on the linguistic motivation for regarding these features as relevant for the task of analyzing semantic change in the following sections.

2 Background

The LSCDiscovery shared task (D. Zamora-Reina et al., 2022) on predicting semantic change for the Spanish language consisted of two sub-tasks. For the first task - graded discovery - the participants were asked to rank the set of content words (N, V, A) in the lemma vocabulary intersection of C1 and C2 according to their degree of semantic change between C1 to C2. The predictions were scored against the ground truth via Spearman's rank-order correlation coefficient.

For the second sub-task - binary change - participants were be asked to classify a pre-selected set of content words (N, V, A) into two classes, 0 for no change and 1 for change. The second sub-task also included two optional sub-tasks on predicting whether the target word undergoing semantic change has gained or lost senses, also formulated as a binary classification problem. Submissions were graded using precision, recall and F1-score.

The data consisted of two corpora of texts in the Spanish language: *old corpus*, created using different sources freely available from Project Gutenberg (containing texts published between 1810 - 1906), and *modern corpus*, created using different sources available from the OPUS project (with texts published between 1994 - 2020).

We participated in the LSCDiscovery shared task on semantic change in the Spanish language with submissions in both main sub-tasks: graded discovery and binary change, as well as the optional tasks on sense gain and sense loss. For all tasks we experimented with approaches based on distances in word embedding spaces combined with hand-crafted linguistic features.

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 187 - 192
May 26-27, 2022 ©2022 Association for Computational Linguistics

3 System Overview

In this section we describe the features and models used to make automatic predictions on the semantic change of target words, for both sub-tasks. We release all the code used for implementing our submissions.[1]

The general method for our submissions in all tasks has consisted of computing, for every given target word, several metrics including embedding distances and linguistic hand-crafted features, and subsequently weighing them as features in a model used to predict the final score. The list of features used consists of the following:

- word embedding cosine similarity scores - 3 different scores according to the different alignment methods (see following section for details)
- word polysemy degree
- number of neological synonyms of the word
- Levenshtein distance to closest English word

In the following subsections we describe in detail both the features and the models used to achieve predictions.

3.1 Word Embedding Distances

The first type of features we used is based on word embedding distances. Following already standard approaches in the study of semantic change based on diachronic corpora, we trained word embeddings separately on the two provided corpora, subsequently used an alignment algorithm to obtain a common embedding space, and finally measured the cosine-distance between each target word's representation in the two embedding spaces, as a proxy for the degree of its semantic shift between the two periods represented in the corpora.

The embedding algorithm we used is word2vec (Mikolov et al., 2013), trained with default parameters in the gensim library. We trained two separate models using the same settings on the tokenized versions of the corpora (non-lemmatized). We then aligned the obtained embedding spaces using three different approaches based on (Artetxe et al., 2016, 2017, 2018a,b), using the open-source code provided by the authors[2]: supervised alignment using a seed word dictionary and a linear mapping method, semi-supervised alignment, optimized for using a small seed word dictionary, and unsupervised alignment based on adversarial training.

We chose to include the semi-supervised and unsupervised approach because of the small list of seed words used (which we assumed could not guarantee a high-quality aligned embedding space using the supervised method). As seed words for the supervised and semi-supervised settings we used the same list of function words in Spanish derived from the NLTK[3] library, considering the ones that also occur in the given corpora.

For all sub-tasks and systems submitted, we used the aligned embedding spaces produced with the method above. From a computational performance perspective, the most costly process was alignment, with the other steps completing in negligible time on a GPU machine (using the default GPUs made available on the Google Colaboratory[4] platform): from seconds for training the supervised models to minutes for training the embedding spaces. For the alignment stage, we ran the algorithms on a CPU machine with an 8-core i7 processor. The supervised alignment completed in approximately 5 minutes, while the semi-supervised and unsupervised methods completed in 5 to 7 hours each. The training phase for building and aligning the embeddings models was the most costly from this perspective, while the actual inference computed for the sample of 4,000 target words was negligible in comparison (consisting only of retrieving cosine distance scores from the embeddings spaces and combining it with linguistic features scores).

Model	Correlation
LinReg with cosine-dist and ling. feat.	0.282
Manual weighting cosine-dist and ling. feat.	(-)0.325
Baseline1	0.092
Baseline2	0.543

Table 1: Results for graded discovery task

3.2 Linguistic Features

Word Polysemy For each word, we computed its polysemy degree by counting the number of synsets it occurs in in WordNet(Miller, 1995), specifically in Open Multilingual WordNet(Bond and Foster, 2013). The degree of polysemy is measured simply

[1] https://github.com/ananana/LSCDiscovery-cotohili

[2] https://github.com/artetxem/vecmap

[3] http://nltk.org/book/

[4] https://colab.research.google.com/

Model	F1	Precision	Recall
Manual weighting of cosine-dist and ling. feat.	0.636	0.353	0.750
DecisionTree with cosine-dist and ling. feat.	0.4	0.143	0.211
Baseline1	0.537	0.846	0.393
Baseline2	0.222	0.500	0.143

Table 2: Results for binary change detection

Model	F1	Precision	Recall
Manual weighing of cosine-dist and ling. feat.	0.462	0.316	0.857
DecisionTree with cosine-dist and ling. feat.	0.111	0.071	0.087
Baseline1	-	-	-
Baseline2	0.211	0.400	0.143

Table 3: Results for optional task on sense gain

as the number of synsets obtained (without distinguishing between polysemy and homonymy).

We assume that polysemy (i.e., the coexistence of several possible meanings for one word) is a relevant feature since it has been shown to be statistically correlated with the rate of semantic change in various previous studies (Bréal, 1897; Ullmann, 1963; Magué, 2005). Bréal (1897) and Ullmann (1963) labelled polysemy as the core of meaning change, considering that change occurs when a secondary or connotative meaning replaces the main or denotative one. Ullmann (1963) underlines the role of discontinuity as a "natural diachronic consequence of the polysemic principle", explained in terms of using a word outside of its initial context, until its original meaning is either forgotten by the speakers, or becomes secondary. Magué (2005) defines polysemy as the synchronic manifestation of semantic change. A possible difficulty in the present task is that WordNet cannot make the difference between polysemic words and homonyms (i.e., words that share the same form, but have different origins and, hence, meanings). Nonetheless, the Spanish language has tended, throughout its history, to avoid the homonymic clashes, either by introducing a graphic distinction (e.g. Sp. *gravar* "to charge" < Lat. *gravare*, vs Sp. *grabar* "to record" < Fr. *graver*), either by simply replacing one of the homonyms by an unambiguous lexeme. Therefore, the cases of possible confusion between polysemy and homonymy are found in a small percentage.

Number of Neological Synonyms As a second feature, we considered the number of synonyms the target word has, in particular neologisms. We extract synonyms for a target word using WordNet (considering all possible senses of the word). In order to select only neological synonyms, we assume a synonym is a neologism (literally, a new word) if it does not occur in the old corpus provided in the shared task.

Our hypothesis is that a word with new synonyms may have diverged from its original semantic pattern, as its new lexical rival could have been increasingly regarded as more suitable for the position of the target word. Obeying the tendency of economy of language, it is counterproductive to have two or more words occupying the same position in the structure of the lexicon, therefore one either migrates to a different semantic field, either undergoes, most often, a semantic specialization (e.g. Lat. *vivenda* "living necessities" > Sp. *vivienda* "living place"), a generalization (Lat. *denarius* "an ancient Roman silver coin, worth ten asses" > Sp. *dinero* "money" in general) or a cohyponymic transfer (i.e. a word designating a certain element of a class shifts as a denomination for another element belonging to the same class, e.g. Lat. *pavus* "peacock" > Sp. *pavo* "turkey"). This shift generally affects the former holder of a position in the lexical system, giving way to new candidates.

Levenstein distance to English Words English has exerted, in recent decades, a strong influence on the Romance languages, materialized both in lexical borrowings, and especially in semantic borrowings or calques (Dworkin, 2012).

We assume that the existence of a virtual cognate in English (we understand by "virtual cognates" two or more descendants of the same etymon in different languages, without being inherited in each language; in this investigation, we considered as "virtual cognates" any pair consisting of a Romance borrowing from a Latin word and the English loanword originated from the same Latin word, e.g. Sp. *directo* and Eng. *direct*) with a similar pronunciation (whether sharing the same meaning or not) may be an indicator that the target word could have been influenced by its English correspondent(Uban et al., 2021a). As an example, we could mention the case of Sp. *servidor*, whose significant divergence from its original meaning could also be due to the new acceptation it gained, in computer science, through a calque of Eng. *server* "a computer that provides client stations with access to files and printers as shared resources to a computer network". We retrieve candidate cognate words in English by using the Levenshtein distances from the target word to any English word in the vocabulary, and choosing the closest English word as a potential cognate. We use the Levenshtein distance to this word as a feature in our model. Here are just a few examples of Spanish - English word pairs identified by using the Levenshtein distances, where the influence of the English meaning on the current use of the word in Spanish is significant: Sp. *administración*, originally "act of administering", influenced by Eng. *administration* came to mean as well "Government (of a country)"; Sp. *contemplar*, originally "to see", also received the meaning "to consider" under the influence of Eng. *contemplate*; Sp. *vegetales* "plants" is also used in the acceptation of Eng. *vegetables* "plant or part of the plant used as food"; Sp. *nominar* "to give a name" acquired as well the meaning of Eng. *nominate* "propose as a candidate for elections or for an award", etc.

3.3 Linguistically-Informed Weighting of Features

For one of our solutions submitted to the second sub-task we attempt to combine the selected features by manually assigning weights to each feature, using expert judgements from linguists specialized in Romance languages and in historical semantics.

Table 4 shows the weights we assigned to each feature. We chose the highest weights to the word embeddings feature, giving more importance to the ones obtained with the supervised alignment approach. For the linguistic features, we considered word polysemy and number of neological synonyms. The range of possible values for these features contains higher numbers than the embedding cosine distances, with comparable ranges between the two linguistic features (natural numbers with no upper limit in theory), which is why we assign lower weights for the linguistic features. We consider polysemy as more important than number of synonyms (considering the theoretical justifications presented above). Since the third linguistic feature, designed to measure the closeness to an English cognate (approximated with Levenshtein distance to the closest English word) is less precise than the other features in the way it is measured, and since its effect on language change can be more complex, it was difficult to decide on a specific relative weight in this case that could be reliable, so we left this feature out of this solution.

Feature	**Weight**
embeddings-cosine-unsupervised	0.1
embeddings-cosine-supervised	0.4
embeddings-cosine-semi-supervised	0.1
nr-neo-synonyms	0.02
wordnet-polysemy	0.05

Table 4: Weights for the different features used, manually assigned with the assistance of linguistic experts

While we did not submit results using manual weighting for the first sub-task on graded discovery, we did incorporate them in our submission for the second sub-task which included an optional task on graded discovery. Due to an error when computing the results, we reported the opposite score to the one generated by the model (with a negative sign), leading to a negative rank correlation with the ground truth. We suggest that, disregarding this error, the results can be considered with an opposite sign, leading to a positive correlation.

For binarizing the results, we used a threshold equal to the median score on the full set of target words.

3.4 Supervised Learning of Feature Weights

As a second solution, we learn the relative weights of each of the features considered using a supervised approach by training a simple model on a very small number of annotated examples. As training data, we used the examples and scores provided by the organizers[5] containing a list of 20 target words along with semantic shift scores.

For sub-task 1 (graded discovery) we used a linear regression model, trained to predict the semantic shift degree on the small set of annotated examples.

For sub-task 2 along with the optional subtasks on binary change, we trained a decision tree model to predict binary labels. We binarized the continuous labels in the annotated examples by setting a threshold equal to the median value of semantic shift on the dataset: any score below this threshold was considered a negative label, and any score above it a positive label.

We additionally analyzed the weights learned by the models in order to gain some insights into the importance awarded automatically to each feature. The linear regression model learned the fol-

[5] https://zenodo.org/record/6300105#.YlK2AXVBxhE

lowing weights for the embedding-based cosine scores: 0.35 for the unsupervised alignment space, 0.91 for the supervised space, and 0.34 for the semi-supervised aligned space. For the linguistic features, the model learned a weight of 1 for the neological synonyms feature, 7 for polysemy degree, and 0.27 for the Levenshtein score to English words. We notice that all weights are positive, and interestingly, that their relative importance matches the one considered for setting weights manually based on linguistic motivations.

For predicting decisions on the optional subtasks of sense gain and sense loss, we combined the predictions for binary change with the values of some of the linguistic features considered which could serve as indicators for sense gains or losses, according to the reasons stated before: we consider a word to have lost a sense if it was predicted to have changed its meaning, and it has any neological synonyms, while polysemy is low (less than 2 senses). Any word which was predicted to have changed its meaning and not lost senses was considered to have gained senses.

4 Results

4.1 Task 1: Graded Discovery

We show our results for sub-task 1 in Table 1. We additionally report here the results obtained with the manual weighting system not submitted to the first sub-task, but submitted to the optional graded change task in the second phase. The baselines consisted of: a skip-gram embeddings model with negative sampling, and orthogonal Procrustes for embedding space alignment (baseline 2), and normalized frequency difference.

4.2 Task 2: Binary Change

Results for sub-task 2 are shown in Table 2. We also submitted predictions for the optional task of sense gain, shown in Table 3. We obtained the second place in terms of recall for sense gain. For sense loss, we do not report detailed results since neither of our systems were able to generate correct predictions (obtaining scores of 0.0).

We notice that, in general, the unsupervised approach using manual weighting of features outperformed the supervised approach. This might be due to the very small size of the annotated data, but is also an encouraging result showing the success of incorporating linguistically informed and expert curated measures for predicting semantic change.

5 Conclusions

We have presented our methods and results in participating in the Spanish semantic change shared task. We proposed a system based in part on word embedding distances, which are already the norm in SOTA models for predicting semantic shift (Schlechtweg et al., 2020), and in part on handcrafted linguistic features, chosen based on theoretical linguistic motivation and on empirical evidence of their relevance to semantic change. While we have done minimal experimentation with the parameters and settings used in training word embeddings, and used supervised models trained on very little data, we obtain encouraging results. For the future, we suggest that combining embedding models trained with more fine-tuned parameters optimized for the given task along with features such as the ones described could lead to improved results. We conclude that incorporating linguistically informed features (aside from word frequency) in computational models for predicting semantic change is a valuable and currently under-explored avenue.

Acknowledgements

This research is supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI—UEFISCDI, CoToHiLi project, project number 108, within PNCDI III.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Michel Bréal. 1897. *Essai de sémantique (science des significations).* Slatkine Reprints, Genéve.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. Lscdiscovery: A shared task on semantic change discovery and detection in spanish. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

Steven N Dworkin. 2012. *A history of the Spanish lexicon: A linguistic perspective*. Oxford University Press on Demand.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Jean-Philippe Magué. 2005. *Changements sémantiques et cognition: différentes méthodes pour différentes échelles temporelles*. Ph.D. thesis, Université Lumière-Lyon II.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23.

Ana Uban, Alina Maria Ciobanu, and Liviu P Dinu. 2019. Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166.

Ana-Sabina Uban, Alina Maria Ciobanu, and Liviu P Dinu. 2021a. Cross-lingual laws of semantic change. *Computational approaches to semantic change*, 6:219.

Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Liviu P Dinu, Simona Georgescu, and Laurentiu Zoicas. 2021b. Tracking semantic change in cognate sets for english and romance languages. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 64–74.

Stephen Ullmann. 1963. *The principles of semantics*. Oxford, Glasgow.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *CogSci*.